Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Attention-HAR: Advanced Human Activity Recognition Using a Deep Learning Model with an Integrated Attention Mechanism

Navid Raisi[1], Mahdi Rezaei[*,2] and Behrooz Masoumi[1]

1. Department of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran.
2. Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, UK

Article Info

Article History:
Received 29 January 2025
Revised 18 February 2025
Accepted 15 March 2025

DOI:10.22044/jadm.2025.15658.2683

Keywords:
Human Activity Recognition, Deep Neural Networks, Attention-HAR, Attention Mechanisms, Video-Based Activity Recognition.

**Corresponding author: m.rezaei@leeds.ac.uk (M. Rezaei).*

Abstract

Human Activity Recognition (HAR) using computer vision is an expanding field with diverse applications, including healthcare, transportation, and human-computer interaction. While classical approaches such as Support Vector Machines (SVM), Histogram of Oriented Gradients (HOG), and Hidden Markov Models (HMM) rely on manually extracted features and struggle with complex motion patterns, deep learning-based models (e.g., Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Transformer-based models) have improved performance but still face challenges in handling occlusions, noisy environments, and computational efficiency. This paper introduces Attention-HAR, a novel deep neural network model designed to enhance HAR performance through three key innovations: Conv3DTranspose for spatial upsampling, ConvLSTM2D for capturing spatiotemporal patterns, and a custom attention mechanism that prioritizes critical frames within sequences. Unlike conventional attention mechanisms, our approach dynamically assigns weights to key frames, reducing the impact of redundant frames and enhancing interpretability and computational efficiency. Experimental results on the UCF-101 dataset demonstrate that Attention-HAR outperforms state-of-the-art models, achieving an accuracy of 97.61%, a precision of 97.95%, a recall of 97.49%, an F1-score of 97.64, and an AUC of 99.9%. With only 1.26 million parameters, the model is computationally efficient and well-suited for deployment on lightweight platforms. These findings suggest that integrating temporal-spatial feature learning with attention mechanisms can significantly improve HAR in dynamic and complex environments.

1. Introduction

Humans perform a diverse range of activities that can broadly be categorized into three groups: (a) movements involving only the human body, (b) interactions with objects, and (c) interactions between individuals (Figure 1) [1]

The primary goal of Human Activity Recognition (HAR) is to automatically identify human activities from video or sensor data [2]. Recent advancements in deep neural networks (DNNs) have significantly improved HAR accuracy; however, recognizing complex and nuanced actions remains an active area of research [3]. HAR applications encompass healthcare (e.g., elderly care, patient monitoring), anomaly detection (e.g., suspicious activity detection), smart homes, personal assistants, entertainment, and autonomous vehicles [4, 5]. HAR methods can be categorized into contact-based and vision-based approaches [6]. Contact-based systems require users to make physical contact to interact with commands,

machines, or devices [7]. Contact-based HAR methods rely on sensor data from accelerometers, wearable sensors, or body-mounted devices [8]. While effective, these systems are less common today due to their cost, complexity, and reliance on specialized equipment. Conversely, non-contact or vision-based HAR methods utilize video or image

Figure 1. Human action types: (a) Human only action, (b) human interaction with objects, and (c) human-to-human interaction [1].

data captured by cameras, offering a non-intrusive alternative. Despite privacy concerns, vision-based approaches are widely adopted for their accessibility and ease of deployment [6], [9] Vision-based HAR techniques can be broadly classified into two major categories:

1. Classical Methods: Traditional handcrafted feature-based approaches such as Global Image Structure (GIST), Scale-Invariant Feature Transform (SIFT), and Histograms of Oriented Gradients (HOG) [10] have been widely used in earlier HAR research. These methods relied on manually extracted features and the application of statistical models like Hidden Markov Models (HMMs), Support Vector Machines (SVMs), and k-Nearest Neighbor (k-NN) for classification. Although computationally efficient, these models struggled with complex spatiotemporal patterns and were limited by their dependency on handcrafted features. As environments became more dynamic and noisier, these traditional methods began to fail in handling occlusions, lighting variations, and subtle motion patterns.

2. Deep Learning Methods: Recent advancements in deep learning-based methods have significantly improved HAR by eliminating the reliance on manual feature extraction. Models like Convolutional Neural Networks (CNNs), Multi-Stream Networks, and Hybrid Architectures are designed to automatically extract hierarchical spatiotemporal features from raw video data [11]. These models have shown remarkable success in capturing high-level representations of motion patterns, making them more robust to dynamic environments, occlusions, and noisy data. However, despite their superior performance, many deep learning models still face challenges in capturing long-range temporal dependencies. Traditional CNN-based architectures often fail to model these long-range dependencies effectively and treat all frames equally, making them inefficient in dynamic scenes where subtle motion patterns are crucial. Furthermore, while Transformer-based models offer improved performance in some contexts, they are computationally expensive and impractical for real-time applications, particularly in edge devices. Despite significant advancements, existing HAR methods still face major challenges, including:

Noisy Environments & Occlusions: Many HAR models struggle to differentiate critical frames from irrelevant ones, making them prone to misclassification in dynamic settings.

Limited Spatiotemporal Modeling: Traditional CNN-based architectures fail to capture long-range dependencies, as they treat all frames equally.

Computational Complexity: Transformer-based HAR models achieve high accuracy but require large-scale datasets and substantial computational resources, making them impractical for real-time and edge-based applications.

To overcome these limitations, we propose Attention-HAR, an advanced deep learning model that integrates an attention mechanism within a Conv3DTranspose and ConvLSTM2D architecture. Our key contributions are:

Conv3DTranspose for Spatial Upsampling: Improves spatial feature extraction while reducing information loss.

ConvLSTM2D for Temporal Learning: Captures sequential dependencies in video frames more effectively than standard CNNs.

Custom Attention Mechanism: Dynamically assigns higher importance to key frames, reducing noise and improving classification accuracy.

Unlike Transformer-based architectures, which require significant computational resources, our model achieves a balance between accuracy and efficiency. Attention-HAR reaches 97.61% accuracy on UCF-101 with only 1.26M parameters, making it suitable for deployment on lightweight platforms.

The rest of this paper is structured as follows: Section 2 reviews related works, emphasizing existing limitations. Section 3 details the proposed Attention-HAR model. Section 4 discusses the experimental results, and Section 5 concludes with future research directions.

2. Related Works

2.1. Classical Approaches

Classical approaches HAR rely heavily on handcrafted features and traditional machine learning models. These methods typically involve two stages: preprocessing and feature extraction.

For example, Reddy et al. [12] used the Difference of Wavelet (DoW) and Difference of Gaussian (DoG) filters for spectral and scale-invariant feature extraction, followed by Nearest Neighbor (NN) classification. While effective for datasets like Weizmann and UCF-11, these approaches fail to capture the spatiotemporal complexity of video data, leading to subpar performance compared to deep learning models. Classical methods fail to handle spatiotemporal features effectively. Deep learning addresses this limitation by enabling automatic feature extraction.

2.2. Deep Neural Networks for HAR

2.2.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have become a foundational component in HAR due to their ability to extract spatial features from video frames. Kumar et al. [13] combined a Gaussian Mixture Model (GMM) and Kalman filters to track human motion characteristics, followed by CNNs for feature extraction. Their approach achieved 90.31% accuracy on the UCF-101 dataset but lacked the temporal modeling capabilities essential for complex action sequences. However, while lightweight and efficient, CNN-based architectures lack the temporal modeling needed for advanced activity recognition.

2.2.2. Hybrid Model :CNN-LSTM Architectures

Hybrid architectures that combine CNNs and Recurrent Neural Networks (RNNs), such as LSTMs, address the limitations of purely spatial models by incorporating temporal features. Vrskova et al. [14] proposed a 3D CNN + ConvLSTM framework to enhance activity recognition across datasets like UCF50mini, MOD20, and LoDVP, achieving accuracies of 87.78%, 78.21%, and 93.41%, respectively. However, the model struggled with noise, parameter optimization, and missing data, highlighting the need for more robust frameworks. Sahoo et al. [15] developed a deep two-way LSTM (DBiLSTM) network to learn patterns in sequential image frames. Using the pre-trained network and data augmentation, they effectively prevented overfitting by virtually increasing the size of the training set. Feature extraction from images is performed with a pre-trained Convolutional Neural Network (CNN). The memory performance achieved accuracies of 97.67%, 95.00%, 73.13%, 92.97%, and 69.74% on the KTH, UCF Sports, JHMDB, UCF101, and HMDB51 datasets, respectively, as reported by Sahoo et al. [15]. Kumar Gour and Rai [16] used long-term recurrent convolutional networks (LRCN) and convolutional short-term memory (ConvLSTM) to extract human activity features and achieved good results on the UCF 50 data set compared to the UCF 101. LRCN had unreliable results on the UCF 101 dataset.

2.2.3. Skeleton-Based and Graph-Based Approaches

Skeleton-based methods utilize 2D or 3D joint data to recognize activities: Li et al. [17] introduced Pose Refinement Graph Convolutional Networks to analyze skeletal data, achieving competitive results for robotic applications. Deyzel and Theart [18] proposed a graph-based approach for one-shot skeleton motion detection, demonstrating an accuracy of 87.4% on NTU RGB+D 120. Chen et al. [19] also used GCN to identify skeleton data, but these methods are suitable for pose data. While efficient for specific scenarios, skeleton-based methods rely heavily on accurate pose detection, which can fail under occlusion or noisy conditions.

2.2.4. Transfer Learning Models

Akarsu and Karacali [20] utilized pre-trained models Resnet18, VGG19, and Alexnet for feature extraction followed by an LSTM for classification. The Resnet18 architecture with 71 layers outperformed the other networks because of its many convolutional layers that capture high-level features. Recent research utilizes deep neural networks for action classification and recognition following feature extraction. Alomar and Cai [21] explored transfer learning for HAR using a pretrained TransNet, a 2D CNN model. While achieving learning speed and classification accuracy, they found that models with fewer layers and parameters perform best.

2.2.5. Attention-Based Models

Attention mechanisms have gained prominence in HAR for their ability to prioritize key frames within a sequence. Chen et al. [22] proposed Spurious-3D Residual Attention Networks (S3D RANs), leveraging residual attention modules to enhance spatiotemporal feature extraction. Although their model achieved 93.30% accuracy on UCF-101, its reliance on computationally expensive 3D convolutions limits scalability for real-time applications. Attention-HAR addresses this limitation with a lightweight design, combining Conv3DTranspose and attention mechanisms to reduce overhead. Despite their promise, attention-based models face limitations: Many implementations rely on dense computational resources, reducing their applicability to lightweight systems.

Table 1. Comparison of some recent methods of human activity recognition.

Paper	Method	Strengths	Limitations
Malik et al. [25]	CNN + LSTM	Using the skeleton–based attention mechanism to improve the accuracy of identifying human activities in occlusion conditions	It has less performance against noise and environmental changes and requires sufficient and accurate data for correct identification
Xing et al. [26]	Semi-supervised Video Transformer	Using semi supervised learning for action recognition to reduce reliance on labeled data and increase temporal stability	Trouble with complex occlusions and highly dynamic scenes and check model performance on datasets outside of Kinetics-400
Mao et al. [27]	attention mechanism + DNN	Reducing computational complexity and improving accuracy through focused feature extraction	Limited performance and scalability of the model to specific data sets and generalizability
Hassan et al. [28]	Deep BiLSTM	Increasing the accuracy of activity identification using bipartite LSTM and feature extraction with transfer learning and reducing the need for labeled data	Performance hinges on high-quality data, a substantial number of features, and might be lower in non-standard or noisy conditions.
Genc et al. [29]	CNN (+3) + LSTM	Increasing the accuracy by using the optimized CNN-LSTM model and increasing the processing speed by reducing the computational complexity	It may require more settings and show lower performance in complex data or with many changes in the environment
Uddin et al. [30]	CNN + ConvLSTM + LRCN	Extraction of spatial and temporal features for medical and monitoring applications	Reducing the accuracy of the model in more complex activities and the need to improve the management of data diversity

Existing models often treat attention mechanisms as black boxes, offering limited interpretability.

Recently, Wei & Wang [23] proposed TCN-Attention-HAR, which integrates Temporal Convolutional Networks (TCN) with an attention mechanism to enhance the extraction of time-dependent features in HAR tasks. Unlike conventional CNN-LSTM models, TCN has a flexible receptive field, allowing it to capture long-range temporal dependencies more effectively. The attention mechanism further enhances the model's performance by assigning higher weights to critical frames, thereby reducing noise and improving classification accuracy. Their experimental results on WISDM, PAMAP2, and USC-HAD datasets demonstrated 1.13%–1.83% improvements in accuracy compared to existing state-of-the-art HAR models.

Given the success of TCN-Attention-HAR in modeling spatial-temporal dependencies while maintaining lightweight architecture, this work motivates our approach to designing a more efficient HAR model with a custom attention mechanism. Unlike TCN, which primarily focuses on sequential data processing, our proposed Attention-HAR leverages Conv3DTranspose and ConvLSTM2D layers to further enhance spatial feature extraction while maintaining low computational overhead.

2.2.6. Transformer-Based Approaches

Originally designed for natural language processing, transformers have demonstrated significant potential in HAR. Wensel et al. [24] developed ViTReT, a Vision Transformer integrated with recurrent layers. Although ViT-ReT achieved 94.70% accuracy on UCF-101, its dependence on large-scale datasets and high computational costs limits its practicality for edge-device deployment. In contrast, Attention-HAR overcomes these challenges with lightweight attention mechanisms, delivering superior performance using only 1.26M trainable parameters. In another study, Dass et al. [31] proposed a hybrid model integrating Transformer and ResNet architectures. This model leveraged temporal relationships and spatial feature extraction to achieve superior performance. However, its high computational and memory demands significantly limit scalability, especially for long video sequences. Moreover, optimal performance requires meticulous hyperparameter tuning, and the model struggles with noisy or insufficiently diverse unlabeled data. These challenges render it less suitable for real-time applications and resource-limited environments. Shi and Liu [32] combined CNNs and Transformers to extract robust spatiotemporal features for HAR. By utilizing advanced pose estimation (MoveNet) and pre-trained convolutional features, they achieved accuracies of 83.41% and 87.50% on the UCF 50 and UCF 101 datasets, respectively. Despite its strong performance, this architecture remains computationally expensive due to the combined

use of CNNs and Transformers. Additionally, it struggles to distinguish visually similar actions (e.g., Jump Rope vs. Soccer Juggling) and lacks generalizability, having been tested only on specific datasets. While Transformer-based models have demonstrated promise, their computational demands often hinder real-time applications.

Attention-HAR addresses this gap by integrating attention mechanisms with light weight components, offering a scalable and efficient solution for HAR.

2.3. Challenges in Existing Models

Despite advancements, many HAR models face critical limitations. For instance, Ullah Khan et al. [33] combined CNNs and LSTMs for Kinect V2 data, achieving an accuracy of 90.89%. However, their model struggled with group activity recognition, highlighting a common challenge in scaling HAR systems to scenarios with multiple participants. Despite significant progress, current HAR models face the following challenges: Lack of Focus on Critical Frames: Existing ConvLSTM and CNN-based models treat all frames equally, leading to suboptimal feature extraction; Generalization to Noisy Environments: Many models struggle with noisy or crowded scenes; Computational Efficiency: High-performing models, such as transformers, are often computationally intensive, limiting their applicability in real-time systems. Table 1 shows the strengths and limitations of some human activity recognition methods.

By integrating Conv3DTranspose, ConvLSTM2D, and a custom attention mechanism, Attention-HAR addresses gaps in prior methods, including computational inefficiencies, sensitivity to noise, and challenges in prioritizing critical temporal features. These innovations allow the model to surpass current state-of-the-art methods on UCF-101 while preserving computational efficiency.

3. Proposed Method

This section provides the details of the model structure.

3.1. Overview

This study presents Attention-HAR, an innovative deep neural network (DNN) architecture developed for HAR. Attention-HAR integrates three core components: **Conv3DTranspose** for enhanced spatial upsampling, **ConvLSTM2D** for capturing temporal dependencies, and **A custom attention mechanism** for identifying and prioritizing critical frames in video sequences. These elements work synergistically to address challenges in HAR, such as noisy environments, occlusions, and subtle action variations, resulting in improved spatiotemporal feature extraction and classification accuracy.

3.2. Preprocessing

Preprocessing is a crucial step to ensure consistency and quality of input data. The following operations are performed: **Frame Extraction**: Video sequences are converted into individual frames for further processing; **Resizing**: All frames are resized to 120 × 120 pixels to ensure uniform input dimensions; **Normalization**: Pixel intensities are scaled to a standard range, improving convergence during model training; **Sequence Alignment**: Frames are truncated or zero-padded to achieve fixed sequence lengths, ensuring compatibility with batch processing.

3.3. Model Architecture

The architecture of Attention-HAR, illustrated in Figure 2, consists of three main modules:

3.3.1. Conv3DTranspose for Spatial Upsampling

We employ 3D Convolutional Neural Networks (3D CNNs) [10,14], but in the transposed mode to extract salient features from activity frames that lead to reduced detection loss. The Conv3DTranspose layer expands the inputs' size, height, and width [34]. The transposed convolutions, called deconvolution, use a transformation in the opposite direction of a normal convolution. This means that the output is obtained from the convolution of the input shape by maintaining a deviation and this convolution is a connectivity pattern [35]. Our Conv3D Transpose layer up-samples the input without losing the original pattern with the same-padding kernel transform with 32 filters and kernel size (3x3x3). Increasing the number of Conv3D Transpose layers and filters can improve salient feature detection. This layer mitigates information loss during convolution operations.

In this model, we use the Leaky ReLU activation function. Leaky ReLU is a variant of the ReLU activation layer that assigns non-zero outputs for negative inputs using the function $f(x) = max(\alpha x, x)$, where α is the parameter defined in the range (0, 1). Leaky ReLU activation function helps the model learn complex patterns by allowing a small gradient when inputs are negative [36]. In the next layer, we use MaxPooling3D to select prominent features.

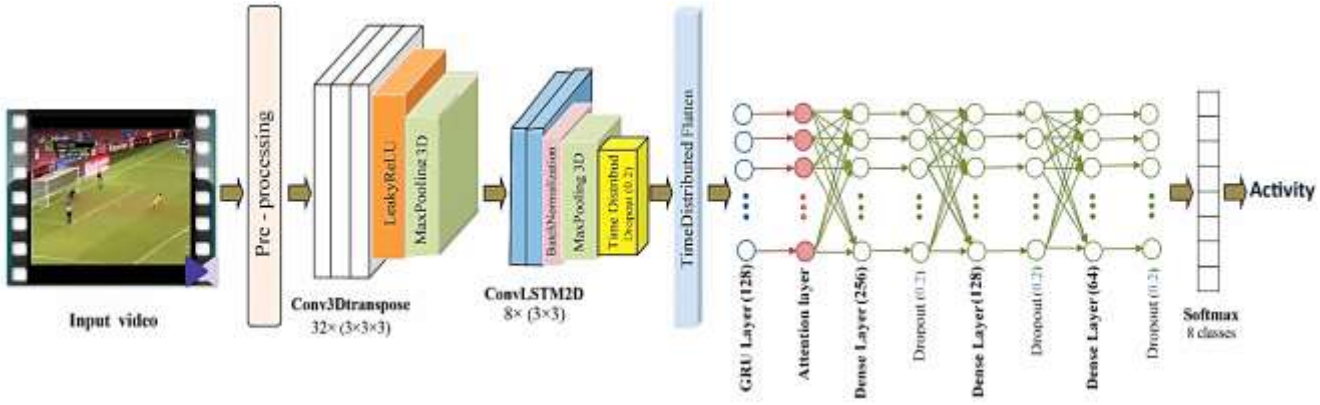

Figure 2. Proposed Model Architecture.

3.3.2. ConvLSTM2D for Temporal Feature Extraction

In the next model layer, we use ConvLSTM2D [37] to extract the sequence of features and reduce the parameters. The ConvLSTM neural network [38] combines a long short-term memory network (LSTM) with a convolutional neural network (CNN). ConvLSTM uses the memory capability of the LSTM network to perform convolution operations on transitions between states and inputs. A ConvLSTM can record the display states of objects and their slow and fast movements using transitions. This layer is particularly effective for video data as it preserves spatial hierarchies [38]. The ConvLSTM2D layers effectively capture spatiotemporal dependencies, thereby improving recognition accuracy. This is particularly useful for detecting subtle temporal patterns, such as hand gestures or object interactions.

The ConvLSTM key equations are derived from the LSTM equations of the convolution pair, as follows:

$$i_t = \sigma(W_{xi} * Xt + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (2)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3)$$

$$O_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \quad (4)$$

$$H_t = O_t \circ \tanh(C_t) \quad (5)$$

The operator '*' represents standard convolution, while '∘' denotes the Hadamard (element-wise) product. The kernels are represented by W. Cell inputs are denoted as X_t, cell states as C_t, hidden states as H_t, and gates as it, f_t, O_t. The sigmoid function is represented by σ [14, 39]. In the proposed Attention-HAR model, we apply MaxPooling3D to the output data of ConvLSTM2D to sample the changes in the position of the features in the input image. Batch Normalization follows this layer to stabilize learning by normalizing outputs. We also added a Time-Distributed layer with a Dropout rate of 0.2 that can receive multiple frames like an overlay layer.

Finally, the multidimensional input is transformed into a one-dimensional format by the convolutional component of the Time Distributed Flatten layer, commonly employed to transition from the convolutional layer to the fully connected layer. Once each frame is flattened using the Time Distributed layer, a gated recurrent unit (GRU) layer further processes the sequence. We use the proposed single-layer GRU [40] model with 128 internal units. GRU reduces the vanishing gradient problem and has the same input and output structure as RNN. Compared to RNN, a GRU with fewer components and state gates can hold critical information longer. Using the GRU layer in this part of the model leads to a greater focus on applying the sequence of features extracted from each activity to the network with fewer parameters, which increases the efficiency of Attention-HAR while preserving critical temporal information. The GRU transmits the output y_t and hidden state h_t to the following node by receiving the current input X_t and hidden state h_{t-1} from the previous node. The Hadamard product is represented by the operator "∘" [41].

$$h_t = (1-z) \circ h_{t-1} + z \circ \tilde{h}_t \quad (6)$$

3.3.3. Custom Attention Mechanism

To improve detection accuracy in the proposed model, we incorporate the attention mechanism [42, 43, 44]. This mechanism allows the model to dynamically assign higher importance to key frames within a sequence, improving

interpretability and classification performance. Unlike conventional models that treat all frames equally, the attention layer identifies and amplifies the most relevant temporal features, making the model more robust in activity recognition tasks. To use attention over the GRU output, we need the GRU to return the entire sequence, not just the last hidden state. This ensures that attention can be applied across all time steps. Incorporating an attention mechanism into the model after the GRU layer allows it to focus on the most relevant parts of the sequence. Here, we implement a simple attention layer using a custom Keras layer. Using the GRU's output, this layer computes attention weights, applies them to generate a context vector, and forwards this vector to the fully connected layers. The model's Attention Layer emphasizes significant frames within the sequence, improving classification accuracy. This is achieved by assigning weights to each timestep's output from the GRU layer, highlighting significant frames. The operation of the attention mechanism [45] is as follows. The custom attention mechanism assigns weights to individual frames, improving interpretability by highlighting which frames contribute most to the model's predictions. This provides clearer insights into the decision-making process, making Attention-HAR well-suited for applications that demand transparency. The key steps in the attention computation are:

1. Score Calculation: Each hidden h_t generated by the GRU layer is assigned a relevance score that indicates its importance in constructing the final context vector, which summarizes the entire input sequence for classification.

2. Weight Normalization: A SoftMax function ensures that the scores sum to 1, producing attention weights.

3. Context Vector Computation: The final context vector is obtained as a weighted sum of all times.

Given an input sequence $X = \{x_1, x_2, ..., x_T\}$, the GRU layer processes these sequences to generate hidden states $H = \{h_1, h_2, ..., h_T\}$ where $h_t \in \mathbf{R}^d$ represents the hidden representation at time step t.

Step 1: Attention Score Computation: Each hidden state h_t is transformed into an intermediate attention score u_t using a trainable weight matrix W and bias vector b :

$$u_t = \tanh(W.h_t + b) \tag{7}$$

Where $W \in \mathbf{R}^{d \times d}$ is a trainable weight matrix that maps each hidden state to an attention-specific representation. $b \in \mathbf{R}^d$ is a trainable bias vector that introduces additional flexibility in learning attention scores. $u_t \in \mathbf{R}^d$ is the intermediate attention score for each time step.

Step 2: Computing Normalized Attention Weights: To ensure that attention scores are normalized across all time steps, we compute the SoftMax attention weight α_t as:

$$\alpha_t = \frac{\exp(u_t^T.u)}{\sum_{j=1}^{T} \exp(u_t^T.u)} \tag{8}$$

where: $u \in \mathbf{R}^{d \times 1}$ is a trainable attention vector that determines the relevance of each time step. $\alpha_t \in [0,1]$ represents the final weight assigned to h_t. The SoftMax function ensures that all attention weights sum to 1, making them probabilistic.

Step 3: Computing the Context Vector *c*:

The final context vector *c* is obtained as a weighted sum of all hidden states:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{9}$$

where $c \in \mathbb{R}$d represents the attention-weighted feature representation of the entire sequence. This vector is then passed to the fully connected layers for final classification.

The weight matrix W and vector u are trainable parameters initialized randomly and optimized during backpropagation. The model updates these values using gradient descent with the Adam optimizer, ensuring the most relevant frames receive higher attention scores. This dynamic weighting mechanism enhances the model's robustness to noisy or redundant frames.

3.3.4. Dense Layers and Classification

The context vector produced by the attention mechanism is fed into fully connected dense layers with 256, 128, and 64 neurons, each utilizing Leaky ReLU activation. Dropout layers with a 0.2 rate are employed to prevent overfitting. The final layer is a SoftMax classifier that outputs probabilities for each activity class. SoftMax function determines a probability in the range of 0 to1, converting an integer vector into a probability vector. The standard (unit) SoftMax function is represented by $\sigma : \mathbf{R}^k \to (0,1)^k$ where $k \geq 1$ is defined by:

$$\sigma(z)_t = \frac{e^{z_i}}{\sum_{j=1}^{k} e^z j} \tag{10}$$

for i=1,2,3,...,8 and $z = (z_1, z_2, ..., z_8) \in \mathbf{R}^k$

3.4. Model Parameters and Complexity

Table 2 outlines the model's layers, detailing the number of trainable and non-trainable parameters. Attention-HAR's architecture is optimized to balance accuracy and computational efficiency. With roughly 1.26 million trainable parameters, the model is lightweight compared to larger architectures like transformers, making it ideal for real-time applications and resource-limited environments.

Table 2. Layers and Number of Parameters in the Proposed Attention-HAR Model.

Layer (type)	Output shape	# of Parameters
Conv3DTranspose	(None, 8, 120, 120,32)	2,624
LeakyReLU	(None, 8, 120, 120, 32)	0
MaxPooling3D	(None, 3, 40, 40, 32)	3
ConvLSTM2D	(None, 3, 38, 38, 8)	11,552
BatchNormalization	(None, 3, 38, 38, 8)	32
MaxPooling3D	(None, 3, 19, 19, 8)	0
TimeDistributed	(None, 3, 19, 19, 8)	0
TimeDistributed-1	(None, 3, 2888)	0
GRU	(None,3, 128)	1,158,912
AttentionLayer	(None, 128)	16,640
Dense	(None, 256)	33,024
Dropout-1	(None, 256)	0
Dense-1	(None, 128)	32,896
Dropout-2	(None, 128)	0
Dense-2	(None, 64)	8,256
Dropout-3	(None, 64)	0
Dense-3	(None, 8)	520

Total params: 1,264,456 (4.82 MB)
Trainable params: 1,264,440 (4.82 MB) Non-trainable params: 16 (64.00 Byte)

4. Experimental Results

4.1. Dataset and Evaluation Metrics

The UCF-101 dataset, a benchmark for HAR, consists of 13,320 video clips across 101 activity classes. For this study, we selected eight representative classes from the UCF-101 dataset: Playing Guitar, Playing Dhol, Typing, Soccer Penalty, Wall Pushups, Surfing, Bowling, and Playing Cello. The selection was made to ensure diversity in movement types (sports, musical performances, and general activities) and to address recognition challenges associated with different actions. Figure 3 shows these 8 activities of the UCF101 video dataset. To ensure robust evaluation:

1. The data was divided into 75% for training and 25% for testing
2. k-fold cross-validation $k = 5$ was applied to validate the model across different splits and prevent overfitting.

The data is provided to the model as a training tensor, utilizing the Adam optimization function with a minimum learning rate threshold of 1×10^{-6}. It has been used to reduce the learning rate by observing changes in the validation error. If the validation error does not improve, the learning rate is halved. This adaptive strategy allows the model to converge more smoothly. We use "categorical cross-entropy" loss as this is a multi-class classification task. To manage the number of epochs based on the lowest loss during the compile stage, the "Early Stopping" method is employed. This approach helps prevent overfitting and reduces training time. The Early Stopping callback halts training if the validation loss fails to improve for 10 consecutive epochs, restoring the best weights from the training process.

Figure 3. Examples of video frames from the UCF-101 dataset, showcasing the diversity of activity classes (e.g., Playing Guitar, Typing). These samples underscore the challenges in HAR, including variations in motion patterns, camera angles, and environmental conditions, all of which are effectively handled by Attention-HAR.

The model is assessed based on performance evaluation and validation criteria. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) form the basis of these criteria. Accuracy is calculated as the percentage of correctly identified predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

In addition to the standard categorical accuracy, we employ Top-K Categorical Accuracy with $k = 5$. This metric assesses whether the true label is within the top 5 predicted classes for each input, making it especially relevant for tasks with multiple plausible categories or for ensuring robustness when precise classification is challenging. The "Top-K Accuracy" function is formally defined as follows:

$$Top-K \text{ Accuracy} = \frac{1}{N}\sum_{i=0}^{N} 1(TrueLabel_i \in Top-K \text{ Predictions}_i)$$

where N is the total number of samples, and 1 is an indicator function that is 1 if the true label appears in the top K predictions and 0 otherwise.

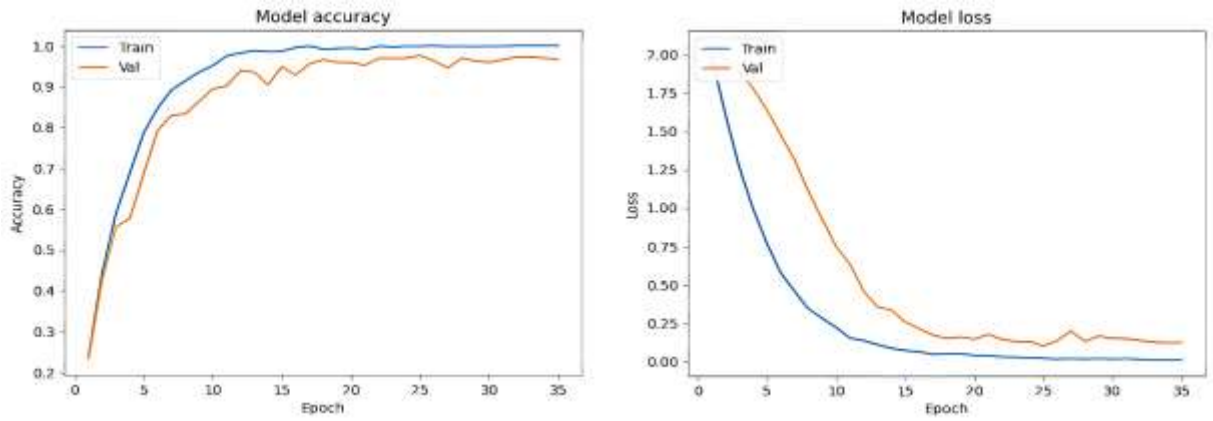

Figure 4. The graph of changes in accuracy and loss of the Attention-HAR model test based on train and validation data.

Table 3 presents the top 5 classification accuracies, the detection accuracy of the proposed method for testing 8 classes on the UCF-101 dataset and training data, along with the area under the curve (AUC) of the receiver operating characteristic (ROC), which evaluates the true positive rate (TPR) against the false positive rate (FPR). To ensure the statistical reliability of the results, we report the standard error (SE) for accuracy and Top-5 Accuracy in this Table. SE is calculated as follows:

$$SE = \frac{\sigma}{\sqrt{N}} \tag{12}$$

Where σ is the standard deviation of the respective metric over multiple models runs, and N is the number of test samples in the respective class.

The training and validation accuracy and the corresponding loss values are shown in Figure 4. The close alignment between training and testing performance indicates minimal overfitting, underscoring the model's generalization capability.

Table 3. Measured KPIs for the Attention-HAR model. on the UCF-101 dataset

Metric	Train (%)	Test (%)	SE (Test) [%]
Accuracy	100.00	97.61	0.89
Top-5 Accuracy	100.00	99.66	0.34
Loss	0.0109	0.1217	-
ROC-AUC	-	99.90	0.19

4.2. Per-Class Performance

The correct detection of positive samples is achieved using the precision formula, which is considered one of these criteria and is defined as follows [9]:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{13}$$

Recall, also known as sensitivity, evaluates the model's ability to identify all true positives. It is calculated using the formula.

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{14}$$

Precision and recall are balanced by the F1-score, which is calculated as follows.

$$F1score = 2 \times \frac{\Pr ecision \times \mathrm{Re}\,call}{\Pr ecision + \mathrm{Re}\,call} \tag{15}$$

Table 4 presents the per-class performance, highlighting Attention-HAR's ability to distinguish between diverse activities. Given the relatively balanced dataset, accuracy remains a reliable metric. To ensure the statistical robustness of these results, we report the Standard Error (SE) for each metric.

Table 4. Detailed Per-Class Results with Standard Error.

Activity	Samples per class	precision	Recall	F1-score	SE precision [%]
PlayingGuitar	40	0.9750	0.9750	0.9750	2.47
PlayingDhol	41	0.9111	1.0000	0.9534	4.44
Typing	34	1.0000	0.8823	0.9375	0.00
SoccerPenalty	34	1.0000	1.0000	1.0000	0.00
WallPushups	33	1.0000	1.0000	1.0000	0.00
Surfing	31	1.0000	0.9677	0.9836	0.00
Bowling	39	0.9743	0.9743	0.9743	2.53
PlayingCello	41	0.9761	1.0000	0.9879	2.39
Macro avg	293	0.9795	0.9749	0.9764	-
Weighted avg	293	0.9774	0.9761	0.9759	-

4.3. Confusion Matrix Analysis

The confusion matrix (Figure 5) highlights occasional misclassifications between Typing and Playing Dhol, likely due to overlapping motion features. However, the consistently high precision across other activities reinforces the robustness and

adaptability of Attention-HAR in diverse scenarios.

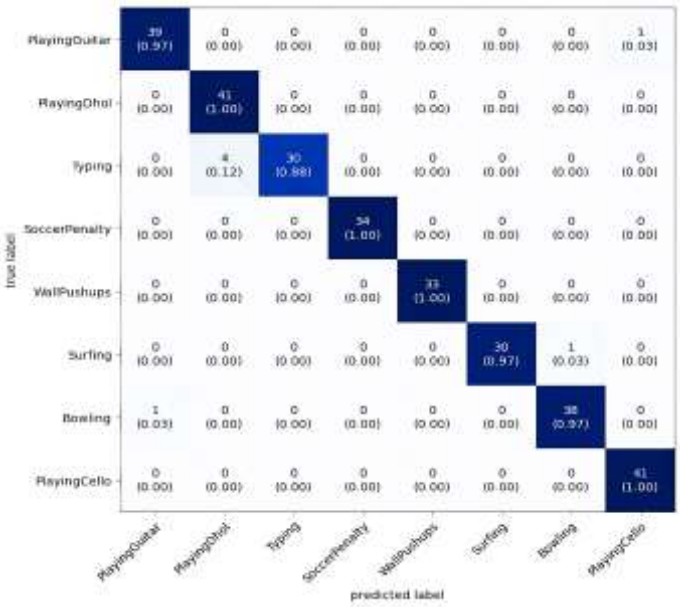

Figure 5. Confusion matrix depicting classification performance across eight UCF101 activity classes, highlighting robust detection for most activities and occasional overlap between 'Typing' and 'Playing Dhol'.

4.4. Comparison with State-of-the-Art Models

Table 5 demonstrates Attention-HAR's superiority over existing models in both accuracy and computational efficiency. By integrating Conv3DTranspose, ConvLSTM2D, and attention mechanisms, Attention-HAR delivers enhanced accuracy while preserving computational efficiency. These results demonstrate Attention-HAR's superiority in balancing accuracy, scalability, and robustness in dynamic environments.

Table 5. Comparison of Attention-HAR's performance against current top models.

Methods	Year	Dataset	Accuracy (%)
Hybrid deep learning [46]	2020	UCF101	89.3
R [2-1] D [47]	2021	UCF101	78.70
CNN-LSTM [48]	2022	UCF101	79.21
RNN [49]	2023	UCF101	90.74
CNN + transform [32]	2024	UCF101	87.50
ViT-ReT [24]	2023	UCF101	94.7
AVSO + AST-VLAD [48]	2024	UCF101	96.0
CNN-CAM+AE [50]	2024	UCF101	97.16
Attention-HAR	**2025**	**UCF101**	**97.61**

5. Conclusion

This study introduces Attention-HAR, a cutting-edge deep neural network model for HAR that effectively captures spatiotemporal dependencies while maintaining computational efficiency. Leveraging Conv3DTranspose, ConvLSTM2D, and a custom attention mechanism, the model enhances spatial feature extraction, captures spatiotemporal dependencies, and prioritizes relevant frames for improved accuracy and interpretability.

Evaluated on the UCF-101 dataset, Attention-HAR achieves a state-of-the-art accuracy of 97.61%, demonstrating its effectiveness in distinguishing diverse activities. With only 1.26 million trainable parameters, the model also maintains a lightweight architecture, making it suitable for deployment on resource-constrained platforms. The robust performance and computational efficiency of the model, position it for deployment in various fields, including healthcare, autonomous vehicles, and surveillance systems. However, given that UCF-101 lacks real-world challenges such as occlusions and noisy environments, development of new datasets followed by further exploration and evaluation can be considered as possible future works, to improve generalizability.

Future research can also focus on optimizing real time deployment on edge devices, extending to group activity recognition and integrating privacy-preserving frameworks like federated learning. By bridging research and practical applications, Attention-HAR represents a significant advancement in HAR technology.

References

[1] Z. Malik and M. I. Bin Shapiai, "Human action interpretation using convolutional neural network: a survey," *Mach Vis Appl*, vol. 33, no. 3, 2022.

[2] P. Khaire and P. Kumar, "Deep learning and RGB-D based human action, human–human and human–

object interaction recognition: A survey," *J Vis Commun Image Represent*, vol. 86, no. May, p. 103531, 2022.

[3] S. A. Khowaja and S.-L. Lee, "Semantic Image Networks for Human Action Recognition," [Online]. http://arxiv.org/abs/1901.06792. 2019.

[4] H. B. Zhang, Y.X Zhang, B. Zhong, Q. Lei, L. Yang, J.X. Du, and D.S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors (Switzerland)*, vol. 19, no. 5, pp. 1–20, 2019.

[5] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 6, pp.865-878, 2012.

[6] X. Wenkai and E. J. Lee, "Continuous gesture trajectory recognition system based on computer vision," *Applied Mathematics and Information Sciences*, vol. 6, no. 2 SUPPL., pp. 339–346, 2012.

[7] B. Paulson, D. Cummings, and T. Hammond, "Object interaction detection using hand posture cues in an office setting," *Int J Hum Comput Stud*, vol. 69, no. 1, pp. 19–29, 2011.

[8] M. Hossein Shayesteh, B. Shahrokhzadeh, and B. Masoumi, "Game Theory Solutions in Sensor-Based Human Activity Recognition: A Review," Journal of AI and Data Mining, vol. 11, no. 2, doi: 10.22044/jadm.2023.12538.2407. 2023.

[9] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimed Tools Appl*, vol. 79, no. 41–42, pp. 30509–30555, 2020.

[10] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, "Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN," *Sci Afr*, vol. 21, 2023.

[11] Y. Kong and Y. Fu, *Human Action Recognition and Prediction: A Survey*, vol. 130, no. 5. Springer US, 2022.

[12] G.V. Reddy, K. Deepika, L. Malliga, D. Hemanand, C. Senthilkumar, S. Gopalakrishnan, and Y. Farhaoui, "Human Action Recognition Using Difference of Gaussian and Difference of Wavelet," *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 336–346, 2023.

[13] B. S. Kumar, S. V. Raju, and H. V. Reddy, "Human Action Recognition Using a Novel Deep Learning Approach," *IOP Conf Ser Mater Sci Eng*, vol. 1042, no. 1, p. 012031, 2021.

[14] R. Vrskova, P. Kamencay, R. Hudec, and P. Sykora, "A New Deep-Learning Method for Human Activity Recognition," *Sensors*, vol. 23, no. 5, 2023.

[15] S. P. Sahoo, S. Ari, K. Mahapatra, and S. P. Mohanty, "HAR-Depth: A Novel Framework for Human Action Recognition Using Sequential Learning and Depth Estimated History Images," *IEEE Trans Emerg Top Comput Intell*, vol. 5, no. 5, pp. 813–825, 2021.

[16] R. K. Gour and D. Rai, "Unveiling Human Actions: Vision-Based Activity Recognition Using ConvLSTM and LRCN Models," in *2024 OPJU International Technology Conference on Smart Computing for Innovation and Advancement in Industry 4.0, OTCON 2024*, Institute of Electrical and Electronics Engineers Inc., 2024.

[17] S. Li, J. Yi, Y. A. Farha, and J. Gall, "Pose Refinement Graph Convolutional Network for Skeleton-based Action Recognition," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.07367.

[18] M. Deyzel and R. P. Theart, "One-shot skeleton-based action recognition on strength and conditioning exercises," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5167–77. [Online]. https://github.com/michaeldeyzel/SU-EMD. 2023.

[19] H. Chen, Y. Jiang, and H. Ko, "Pose-Guided Graph Convolutional Networks for Skeleton-Based Action Recognition," *IEEE Access*, vol. PP, p. 1, 2022.

[20] E. Akarsu and T. Karacalı, "Video Classification Results with Artificial Intelligence and Machine Learning," *International Journal of Innovative Research and Reviews (INJIRR)*, vol. 7, pp. 22–26, [Online].http://www.injirr.com/article/view/194. 2023.

[21] K. Alomar and X. Cai, "TransNet: A Transfer Learning-Based Network for Human Action Recognition," 2023, [Online]. Available: http://arxiv.org/abs/2309.06951.

[22] B. Chen, H. Tang, Z. Zhang, G. Tong, and B. Li, "Video-based action recognition using spurious-3D residual attention networks," *IET Image Process*, vol. 16, no. 11, pp. 3097–3111, 2022.

[23] X. Wei and Z. Wang, "TCN-attention-HAR: human activity recognition based on attention mechanism time convolutional network," *Sci Rep*, vol. 14, no. 1, Dec. 2024.

[24] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos," *IEEE Access*, vol. 11, no. June, pp. 72227–72249, 2023.

[25] N. ur R. Malik, S. A. R. Abu-Bakar, U. U. Sheikh, A. Channa, and N. Popescu, "Cascading Pose Features with CNN-LSTM for Multiview Human Action Recognition," *Signals*, vol. 4, no. 1, pp. 40–55, Mar. 2023.

[26] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "SVFormer: Semi-supervised Video Transformer for Action Recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18816–26. 2023.

[27] Y. and J. X. and H. Z. and P. Y. Mao Keming and Xiao, "KS-FuseNet: An Efficient Action Recognition Method Based on Keyframe Selection and Feature

Fusion," in *Pattern Recognition and Computer Vision*, M.-M. and H. R. and U. K. and S. W. and Z. H. and Z. J. and L. C.-L. Lin Zhouchen and Cheng, Ed., Singapore: Springer Nature Singapore, pp. 540–553. 2025.

[28] N. Hassan, A. S. M. Miah, and J. Shin, "A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition," *Applied Sciences (Switzerland)*, vol. 14, no. 2, Jan. 2024.

[29] E. Genc, M. E. Yildirim, and Y. B. Salman, "Human activity recognition with fine-tuned CNN-LSTM," *Journal of Electrical Engineering*, vol. 75, no. 1, pp. 8–13, Feb. 2024.

[30] S. Uddin, T. Nawaz, J. Ferryman, N. Rashid, M. Asaduzzaman, and R. Nawaz, "Skeletal Keypoint-Based Transformer Model for Human Action Recognition in Aerial Videos," *IEEE Access*, vol. 12, no. January, pp. 11095–11103, 2024.

[31] S. D. S. Dass, H. B. Barua, G. Krishnasamy, R. Paramesran, and R. C.-W. Phan, "ActNetFormer: Transformer-ResNet Hybrid Method for Semi-Supervised Action Recognition in Videos," [Online]. Available: http://arxiv.org/abs/2404.06243. 2024.

[32] C. Shi and S. Liu, "Human action recognition with transformer based on convolutional features," *Intelligent Decision Technologies*, vol. 18, no. 2, pp. 881–896, May 2024.

[33] I. U. Khan, S. Afzal, and J. W. Lee, "Human activity recognition via hybrid deep learning-based model," *Sensors*, vol. 22, no. 1, 2022.

[34] S. Tomassini, H. Anbar, A. Sbrollini, M. Morettini, M. H. D. J. Mortada, and L. Burattini, "A Double-Stage 3D U-Net for On-Cloud Brain Extraction and Multi-Structure Segmentation from 7T MR Volumes," 2023.

[35] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," [Online]. Available: http://arxiv.org/abs/1603.07285. 2016.

[36] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *International Conference on Digital Signal Processing, DSP*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017.

[37] P. Dasari, L. Zhang, Y. Yu, H. Huang, and R. Gao, "Human Action Recognition Using Hybrid Deep Evolving Neural Networks," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2022-July, 2022.

[38] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.04214.

[39] Y.-L. Chang, *N.B. Tatini, T.H. Chen, M.C. Wu, J.H. Chuah, Y.T. Chen, and L. Chang*, "Convlstm Neural Network for Rice Field Classification from Sentinel-1A Sar Images," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5047–5050. 2022.

[40] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," [Online]. Available: http://arxiv.org/abs/1409.1259. 2014.

[41] L. Lu and K. A. I. Cao, "A Multichannel CNN-GRU Model for Human Activity Recognition," *IEEE Access*, vol. 10, no. June, pp. 66797–66810, 2022.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN. Gomez, Ł. Kaiser, I. Polosukhin. "Attention Is All You Need," *Advances in neural information processing systems, vol* 30, 2017.

[43] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), pp.3279-3298. 2022.

[44] M. H. Guo, *T.X. Xu, J.J. Liu, Z.N. Liu, P.T. Jiang, T.J. Mu, S.H. Zhang, R.R. Martin, M.M. Cheng, and S.M. Hu*, "Attention mechanisms in computer vision: A survey," *Tsinghua University*. 2022.

[45] D. Kumari and R. S. Anand, "Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism," *Electronics (Switzerland)*, vol. 13, no. 7, Apr. 2024.

[46] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447–453, 2020.

[47] Pan, Tian, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. "Videomoco: Contrastive video representation learning with temporally adversarial examples." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11205-11214. 2021.

[48] D. Kumar, M., Rana, A., Ankita, Yadav, A. K., & Yadav, "Human Activity Recognition in Videos Using Deep Learning," in *In International Conference on Soft Computing and its Engineering Applications*, pp. 288–299. 2022.

[49] A. Alavigharahbagh, V. Hajihashemi, and J. J. M. Machado, "Deep Learning Approach for Human Action Recognition Using a Time Saliency Map Based on Motion Features Considering Camera Movement and Shot in Video Image Sequences,", *Information 14*(11), p.616, 2023.

[50] E. Dastbaravardeh, S. Askarpour, M. Saberi Anari, and K. Rezaee, "Channel Attention-Based Approach with Autoencoder Network for Human Action Recognition in Low-Resolution Frames," *International Journal of Intelligent Systems*, no 1, 1052344, 2024.

Attention-HAR: بازشناسی پیشرفته فعالیت انسانی با استفاده از مدل یادگیری عمیق به همراه مکانیزم توجه

نوید رئیسی[1]، مهدی رضائی[2,*] و بهروز معصومی[1]

[1] گروه مهندسی کامپیوتر و فناوری اطلاعات، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران.

[2] موسسه مطالعات حمل و نقل، دانشگاه لیدز، لیدز، انگلستان.

ارسال ۲۰۲۵/۰۱/۲۹؛ بازنگری ۲۰۲۵/۰۲/۱۸؛ پذیرش ۲۰۲۵/۰۳/۱۵.

چکیده:

بازشناسی فعالیت انسانی با استفاده از بینایی کامپیوتر یک زمینه در حال توسعه با کاربردهای متنوع از جمله مراقبت های بهداشتی، حمل و نقل و تعامل انسان و کامپیوتر است. در حالی که رویکردهای کلاسیک مانند ماشین های بردار پشتیبان، هیستوگرام گرادیان های جهت یافته و مدل های پنهان مارکوف بر ویژگی های استخراج شده دستی تکیه دارند و در مواجهه با الگوهای حرکتی پیچیده دچار چالش می شوند، مدل‌های مبتنی بر یادگیری عمیق مانند شبکه های عصبی کانولوشنال، شبکه حافظه بلند کوتاه مدت بازشناسی فعالیت را بهبود می بخشند. با این حال انسداد فعالیت، محیط های شلوغ و پیچیدگی محاسباتی از جمله چالش های بازشناسی فعالیت هستند. این مقاله روشی به نام Attention-HAR را معرفی می‌کند؛ یک مدل شبکه عصبی عمیق جدید که برای بهبود عملکرد بازشناسی فعالیت انسان از طریق سه نوآوری کلیدی طراحی شده است: Conv3DTranspose برای نمونه برداری فضایی، ConvLSTM2D برای گرفتن الگوهای مکانی-زمانی، و در نهایت مکانیزم توجه سفارشی که قاب های برجسته را در توالی ها انتخاب می‌کند. برخلاف مکانیسم های توجه مرسوم، رویکرد ما به صورت پویا وزن هایی را به قاب های کلیدی اختصاص میدهد که تأثیر فریم های اضافی را کاهش و تفسیر پذیری و کارایی محاسباتی را افزایش می‌دهد. نتایج آزمایشات روی مجموعه داده UCF-101 نشان داد که Attention-HAR از آخرین مدل‌های پیشرفته بهتر عمل می‌کند و به دقت ۹۷/۶۱٪، صحت ۹۷٪/۹۵، حساسیت ۹۷/۴۹٪، میانگین هارمونیک دقت و حساسیت ۹۷/۶۴٪ و آستانه اندازه‌گیری ۹۹/۹٪ دست می‌یابد. مدل ارائه شده ما با ۱/۲۶ میلیون پارامتر، از نظر محاسباتی کارآمد و برای اجرا بر روی پلتفرم های سخت افزاری با امکانات پردازش محدود مناسب است. این یافته ها نشان می‌دهد که ادغام یادگیری ویژگی‌های زمانی- مکانی با مکانیسم های توجه می‌تواند به طور قابل توجهی بازشناسی فعالیت انسان را در محیط های پویا و پیچیده بهبود بخشد.

کلمات کلیدی: بازشناسی فعالیت انسانی، شبکه های عصبی عمیق، Attention-HAR ، مکانیسم‌های توجه، بازشناسی فعالیت مبتنی بر ویدئو.